90/07/17

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE

Massively Parallel Network Architectures for Automatic Recognition of Visual Speech Signals

5. FUNDING NUMBERS

AFOSR-86-0246

6. AUTHOR(S)

Terrence J. Sejnowski
Moise Goldstein

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Johns Hopkins University
3400 N. Charles Street
Baltimore, MD 21218

8. PERFORMING ORGANIZATION REPORT NUMBER

AFOSR-TR- 90 0949

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFOSR/NE
Bldg. 410
Bolling AFB, DC 20332

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

2305/B3

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release;
distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

see attached technical report

14. SUBJECT TERMS	15. NUMBER OF PAGES
90 09 20 020	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED		unlimited

FINAL TECHNICAL REPORT

GRANT AFOSR-86-0246

PI: **Terrence J. Sejnowski**

CO-PI: Moise Goldstein

TITLE: **Massively Parallel Network Architectures for Automatic Recognition of Visual Speech Signals**

ADDRESS:

The Johns Hopkins University
Charles and 34th Streets
Baltimore, MD 21218

STARTING DATE: 7/1/86

SUMMARY

The goal of the research performed under this grant was to produce a massively-parallel network architecture that could interpret speech signals from video recordings of human talkers. This report summarizes the results of this project: 1) A corpus of video recordings from two human speakers was analyzed with image processing techniques and used as the data for this study; 2) We demonstrated that a feedforward network could be trained to categorize vowels from these talkers. The performance was comparable to that of the nearest neighbors technique and to trained humans on the same data; (3) We developed a novel approach to sensory fusion by training a network to transform from facial images to short-time spectral amplitude envelopes. This information can be used to increase the signal-to-noise ratio and hence the performance of acoustic speech recognition systems in noisy environments; (4) We explored the use of recurrent networks to perform the same mapping for continuous speech.

The results of this project demonstrate the feasibility of adding a visual speech recognition component to enhance existing speech recognition systems. Such a combined system could be used in noisy environments, such as cockpits, where improved communication is needed. This demonstration of presymbolic fusion of visual and acoustic speech signals is consistent with our current understanding of human speech perception. Further studies are continuing to extend these results from vowels to consonants, and to develop improved networks for combining information streams from multiple sources.

PUBLICATIONS

Yuhas, B. P., Goldstein, M. H., Jr. & Sejnowski, T. J., Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine* , November, 65-71 (1989).

Yuhas, B. P., Goldstein, M. H. Jr., Jenkins, R. E., & Sejnowski, T. J., Combining visual and acoustic speech signals with a neural network improves intelligibility. In: D. Touretzky (Ed.) **Advances in Neural Information Processing Systems**, 2. San Mateo, California: Morgan Kaufmann Publishers, 1990.

Yuhas, B. P., Goldstein, M. H., Jr., Jenkins, R. E. & Sejnowski, T. J., Neural network models of sensory integration for speech. *Proceedings of the IEEE* (in press)

INTRODUCTION

Automatic speech recognizers currently perform poorly in the presence of noise. Humans, on the other hand, often compensate for noise degradation by extracting speech information from alternative sources and then integrating this information with the acoustic signal. Loss of information in the acoustic signal can be compensated for by using information about speech articulation from the movements around the mouth, or by using semantic information conveyed by facial expressions and other gestures. The listener can also use knowledge of linguistic constraints to compensate for ambiguities remaining in the received speech signals.

Efforts have been made to reduce the noise in the acoustic signal, but few have attempted to use additional external information sources. One notable exception is a system built by Eric Petajan (1987) for isolated digit recognition that used vector-quantized binary images of the speaker's mouth. In this system, the acoustic and visual speech information were independently encoded into symbol strings, and a set of rules was used to reconcile conflicting interpretations. The symbolic intermediates were needed to perform the necessary processing and integration in real time on the serial digital computers available. The massively-parallel architecture of artificial neural networks make it feasible to explore subsymbolic alternatives to Petajan's system. The use of high-dimensional representations allows information from several sources to be combined "softly", before being reduced to discrete symbols. In addition, learning algorithms provide a means of training networks to fuse these signals without explicit rules or restrictive *a priori* models.

The approach taken here was to use the visual speech signals to clean up the acoustic signal. Neural networks are trained to estimate the associated acoustic structure from the concurrent visual speech signal. This acoustic estimate was then fused with the noise-degraded acoustic information. By combining the visual and acoustic sources of speech information, we demonstrated that the visual signal can be used to improve the performance of automatic vowel recognition in the presence of noise. This approach does not require categorical preprocessing or explicit rules. The performances of these neural networks compared favorably with human performance and with other pattern-matching and estimation techniques. Our results were based on vowels spoken by single speakers, but this same approach can be extended to multiple speakers and to consonants.

NEURAL NETWORKS

The architecture of artificial neural networks is motivated by the computational style found in biological nervous systems. The key features are a large number of relatively simple nonlinear processing units and a high degree of connectivity between these units. A unit performs a nonlinear transformation on the sum of its inputs to produce an output signal. When this output signal travels across a connection to another unit, the signals is attenuated or amplified by the weight associated with that connection. Computation is performed by the interaction of these units and signals. Rather than having an explicit program, the computation is defined by the properties of the individual units and their interconnects. As architectural abstractions, these models differ from actual neural networks found in the nervous systems. For example, the processing units used in this study simply add their weighted inputs and have a static sigmoidal nonlinear output function, while neurons in real nervous systems have more complex spatiotemporal nonlinearities and are capable of much more complex discriminations. Nevertheless, as an architecture, these networks provide alternative approaches to difficult computational problems.

Feedforward network architectures were used in most of this study. The units in a feedforward network were arranged in layers, with connections only allowed between layers, and only in one direction. The units that receive inputs from outside the network are referred to as input units, and those that are observed from outside the network are output units. The remaining units are referred to as hidden, because they only exchange signals with other parts of the network. The units themselves use a nonlinear sigmoid squashing function to transform the sum of their inputs. The standard multilayered feedforward networks with arbitrary squashing functions are a class of universal approximators. Moreover, any nonlinear mapping can be learned by a network if there are sufficient data to characterize the mapping and if the number of parameters in the network matches the information content of the data.

A modified backpropagation algorithm was used to train feedforward networks (Rumelhart, et al., 1986). The gradient was calculated in the standard manner, but instead of using steepest descent, a conjugate-gradient algorithm was used to update the weights. In addition, the fixed-step size and momentum term associated with backpropagation were replaced with a line-search minimization. Our neural networks were simulated on a MIPS M/120 computer and an ANALOGIC AP5000 array-processor. Because of the conjugate gradient learning algorithm, the time it took to perform one backpropagation step varied depending upon the number of evaluations required in the line-minimization search. For a network with 2559 weights it took the MIPS M/120 approximately 35 msec to perform one evaluation. The number of adjustable weights in a neural network can often exceed the

number of training patterns. In these cases, the networks have too many free parameters and are subject to the problem of overfitting or overlearning the training data. The effects of overlearning can be minimized by increasing the size of the training data set, by reducing the number of hidden units, or by stopping the training before the network has completely converged.

THE SPEECH SIGNALS

The speech signals used were obtained from video recordings of a seated speaker facing a camera under well-lit conditions. The visual and acoustic signals were stored on a laser disc (Bernstein and Eberhardt, 1986) where the individual frames and their corresponding speech segments were indexed. The NTSC video standard was used (30 frames/sec) and each frame had 33 milliseconds (ms) of speech associated with it. Phonemes usually are shortened or dropped altogether during fluent speech, so single video frames often span more than one phoneme. To avoid this problem, we selected speech samples such as stressed vowels in isolated word or consonant-vowel-consonant (CVC) type nonsense syllables that change relatively slowly. In these contexts, the vowels often were steady state over periods of 50 to 100 ms. For a given phoneme, a preliminary list of candidate words was identified from a transcription of the laser disc. Each word was then played acoustically to confirm the suspected pronunciation. A representative frame for the vowel was then isolated by alternately dropping a frame and then listening until the surrounding consonants were removed. The number of frames that remained after this process depended upon the degree to which that particular vowel was stressed. Stressed vowels, for example, can last up to 132 ms or 4 frames, while an unstressed vowel in continuous speech will often not last the full 33 ms of a single frame. The acoustic signals of the remaining frames were digitized and visually examined to ensure that acoustic signal was approximately in steady state. From this set, a single frame was selected only if the periodic wave form appeared relatively stable, neither increasing nor decreasing in amplitude. This paper describes results obtained using data from a single male speaker. A data set was constructed of 108 images of 9 different vowels in 12 sets. The vowels were taken from words and CVCs. Because these words and syllables were spoken deliberately and in isolation, these vowels were isolated easily. Data from a female speaker were also studied.

Preprocessing the images. Instead of searching for an optimal encoding of the input images, we chose a simple representation that seemed to contain the relevant information. A rectangular area-of-interest was automatically defined and centered about the mouth. The image was further reduced to produce an image that could be comfortably handled by our network simulations. Within the rectangle, the average value of each 4 x 4 pixel squares was computed to produce a topographically accurate grey-scale image of 20 x 25 pixels. Rather than attempting to extract special features,

this encoding represented a form that could be obtained easily through an array of analog photoreceptors. Two methods of processing these images of the speaker's mouth were explored. In the first approach, we treated the images categorically and attempted to make hard phonemic decisions directly from the images. Such linguistic identifications can be used to constrain the linguistic interpretation of a noise-degraded acoustic signal. In the second approach, we obtained acoustic information directly from the images by estimating the transfer function of the vocal tract. These independent estimates were then used to constrain the acoustic interpretation of the noise-degraded acoustic signal directly.

The acoustic speech signal emitted from the mouth can be modeled as the response of the vocal-tract filter to a switchable sound source. In a first-order vocal-tract model, the configuration of the articulators (e.g, the mouth opening, the lips, teeth, tongue, velum and glottis) defines the shape of the vocal tract filter, which then determines the filter's frequency response. The resonances of the vocal tract filter appear as peaks in the envelope of the short-term power spectrum of the acoustic signal and are called formants. While some of the articulatory features are often visible (e.g., the lips, teeth and sometimes the tongue), other components of the articulatory system, such as the glottis and velum, are not. Those articulators that are visible tend to modify the acoustic signal in ways that are more susceptible to acoustic distortion than those effects due to the hidden articulators. This complementary structure can be exploited to improve the perception of speech in noise.

CATEGORIZATION

Neural networks were trained to identify the vowel directly from the image. The images were presented across 500 input units, and the output consisted of 9 output units, each representing one of the nine vowels in the data. An input image was correctly categorized when the activation value of the correct vowel unit was larger than all the other output units. The data set of 108 images was split into a test set and a training set of 54 images, each containing a balanced set of vowels. The number of hidden units were varied. A network was trained until the categorization of all 54 images in the training set was perfect. Overtraining was minimized by immediately terminating the training at this point, before the output units were driven to saturation. After the network was trained, it then was tested on the second set of 54 images from the same speaker.

Results. Performance levels were averaged across eight networks having five hidden units, each initialized with different random weights. The networks were trained on 54 patterns. For half of the networks, the training and test sets were reversed. The eight networks trained on the male data obtained an average performance of 76% correct categorizations for the

images in the test set. A nearest neighbor classifier (NN) was constructed using the training data as the set of stored templates and the results compared with the performance of the neural network model. The individual images from the test set were correlated with the stored templates, and the image was classified according to its closest match. The process was repeated again, but with the test and training sets reversed. The NN classifier correctly classified the male data set with an average accuracy of 79%. The performance of the network also compared favorably with two human subjects tested and trained on the same data. After 5 training sessions, the two subjects obtained an average of 70% on the images in the test set, with performances in some follow-up sessions approaching 80%. The types of errors made by the human subjects in these experiments were similar to those made by the network as judged by comparing the confusion matrices.

SUBSYMBOLIC PROCESSING

Summerfield (1987) concluded from psychoacoustic experiments that information from the visual and acoustic modalities must be integrated before phonetic or lexical categorization takes place. The implication made is that the acoustic and visual signal streams share a common representation at their conflux. We have used the vocal tract transfer function as this common representation, and we have shown that networks can be designed for integrating visual and acoustic speech signals using this representation. An estimate of the vocal tract's acoustic characteristics was obtained directly from images of the speaker's mouth. This estimate then served as an independent source of acoustic information and was used to constrain the interpretation of the acoustic signal.

The acoustic speech signal is produced by a source signal that passes through the vocal tract and is emitted from the mouth. For voiced speech, the driving signal is a quasi-periodic pulse train convolved with the glottal wave form. This driving signal's contribution to the short-term acoustic spectrum is a series of harmonics reducing in amplitude by -12 dB per octave. This reduction is partially compensated by the radiation of the acoustical signal from the lips, which produces an effective gain of +6 dB per octave. The spectral envelope of the short-term spectrum that remains after these two effects are removed is the frequency response of the vocal tract filter. The transfer function of the vocal tract can be estimated by measuring the short-term spectral amplitude envelope (STSAE) of the acoustic signal.

There is not enough information in the visual speech signal to completely specify the vocal-tract transfer function. Many different acoustic signals can be produced by vocal tract configurations that correspond to the same visual signal. Thus, the visual signals can provide only a partial description of the vocal tract filter. Nonetheless, it may be possible to obtain a *good* estimate of the vocal tract transfer function if additional constraints are

considered. A feedforward neural network was trained to estimate the STSAE of the acoustic signal directly from the visual signals around the mouth. The estimate of the STSAE was then combined with estimates from acoustic information to improve the signal-to-noise ratio prior to recognition. The same images of the male speaker used in the categorization experiments were used in these experiments. Each video frame had 33 ms of acoustic speech associated with it. The short-term power spectra of the corresponding acoustic data were calculated and the spectral envelopes were obtained using cepstral analysis. Each smoothed envelope was sampled at 32 frequencies to produce a vector of scaler values. These vectors were used to represent the vocal-tract transfer functions corresponding to the images.

Vowels are largely identified by their spectral shape, and in particular by the location of their spectral peaks, or formants. Nevertheless, evaluating the quality of these spectral estimates is significantly more difficult than judging the accuracy of a categorization because the perceptual processes involved in processing the spectral peaks is not a well-understood process. To assay our spectral estimates, a simple vowel recognition system was constructed using a simple feedforward network trained to recognize nine vowels from their STSAEs. The network was trained on 6 examples each of 9 different vowels until its performance was 100% on the training data. This network served as a *perfect* recognizer of the noise-free data and was used to assess the benefit of the visually-estimated spectra when combined with the noise-degraded acoustic spectra.

The vowel recognizer was presented with a STSAE through two channels. The path shown on the right in Fig. 1 was for the information obtained from the acoustic signal, while the path on the left provided spectral estimates obtained independently from the corresponding visual speech signal. The first step was to test the performance of the recognizer when the acoustic spectral envelopes were degraded by noise. Zero-mean random vectors were normalized and added to the training STSAEs to produce signals with signal-to-noise ratios ranging from -12 dB to 24 dB. Noise corrupted vectors were produced at 3 dB intervals from -12 dB to 24 dB. At each noise level, 12 different vectors were produced for each of the STSAE in the set. At each level, the performances of the recognizer on the degraded signals were averaged. The overall performance on the training data fell with decreased signal-to-noise ratios. At -12 dB, the recognizer operated at the chance level, which was 11% with nine vowels in the data set.

The next step was to compensate for the noise degradation by providing an independent estimate of the STSAE from the visual signal, as shown on the left side of Fig. 1. The network on this pathway was trained to estimate the spectral envelopes corresponding to the input images. The data used to train this network were different from the data used to train the recognizer. The noise-degraded acoustic signal was then combined with the

output from the network processing the images to provide a single estimate which is then passed on to the recognizer. The acoustic and visual signals were weighted according to their relative information content to compensate for the degraded performance at the signal-to-noise ratio extremes. At each signal-to-noise ratio was varied and the optimal value was found empirically to vary approximately linearly with the signal-to-noise ratio in dB, from 1 at -12 dB signal-to-noise ratio to 0 at 24 dB. The performance is shown in Fig. 2. Another method of fusing the two spectra was accomplished using a sigma-pi neural network (Rumelhart, et al., 1986). These second-order networks took the estimated STSAE, the noise-degraded acoustic STSAE and a measure of the signal-to-noise ratio as input, and tried to produce a noise-free STSAE as output. In contrast to the simple weighted sum used by first-order units, the units in these second-order networks determine the activation level by summing the weighted product or other units' output. The results from this method were mixed: while the squared-error between the estimated and actual spectra was significantly lower, their categorization was poorer. These results suggest that the vowel recognizer is doing something more complicated than simply making a comparison based upon a squared-error measure. It also raises questions as to the appropriateness of the squared-error measure used for training.

Comparing performance. The quality of the networks' estimates were compared to a combination of two optimal linear-estimation techniques. The first step was to encode the images using a Hotelling or Karhunen-Loeve transform. The images were encoded as five-dimensional vectors defined by the largest principal components of the covariance matrix of the images in the training set. This is an optimal encoding of the images with respect to a least-squared-error (LSE) measure. The next step was to find a mapping from these encoded image vectors to their corresponding short-term spectral amplitude envelopes (STSAEs). The fit was found using a linear least-squares fit. The estimates obtained by this two stage process were significantly poorer in overall mean-squared error. The mean-squared error of the estimates made by the networks were 46% better on the training set and 12% better on the test set. The main objective of this comparison was to show that arbitrary encoding of the images may result in a loss of relevant information. In contrast, the network learning algorithm allows the network to produce its own encoding at the hidden layer based upon relevant features. The activation levels of the five hidden units served to encode the image as did the five-dimensional vectors obtained using principal components. The primary difference is that the encoding found by the network optimized the desired output, while the principal components optimized the LSE reconstruction of the images.

DYNAMICS AND SPEECH

In the work described above, attention was restricted to static visual images, which are inherently ambiguous because they contain incomplete information about the speech articulators. Speech is a dynamic process and the articulators are physical structures that move. At a given moment, their current positions are part of larger dynamic trajectories. These trajectories are constrained by the mechanics of the physical system and by the linguistic rules of the language. Dynamic dependencies could provide additional constraints that can serve to restrict the acoustic interpretation of the visual speech signal. In this section, we outline an approach to introducing dynamic constraints in neural network models. One approach is to have projections from the output units to the input layer (Jordan, 1988) or from hidden units to the input layer (Elman, 1988).

When working with static images, it was possible to use a simple vowel recognizer to test the quality and utility of the acoustic spectra estimated from static images. The success of the vowel recognizer depended on the careful selection of vowels from isolated words or syllables. For continuous speech, however, it is difficult and often impossible to make these definitive identifications of short speech segments taken out of context, so alternative assessments are necessary. Networks with feedback were used to estimate the STSAE from images within a larger context. The performance of the network on continuous speech was evaluated on its ability to preserve the salient features of the spectral sequences, such as the resonances, or formants, of the estimated vocal tract filter. To see how well these formants were identified by the network, the sequences of spectra were arranged in a visual display similar to a spectrogram. The spectrogram shown in Fig. 3 was created from spectra estimated from a sequence of images not in the training set. In this form, we can observe the changes of energy in the different frequency bands as a function of time. Clearly, much of the acoustic structure was being estimated in these sequences. The ultimate test will be to either resynthesize the acoustic speech signal from these estimated acoustic parameters, or to feed the fused spectra into a full-scale speech recognizer.

CONCLUSIONS

Under noisy conditions, speech recognition can be aided by extracting information from the visual speech signals and combining it with residual acoustic information. Two representations for the speech information in the visual signal were studied under this grant, both of which can be combined with information from the acoustic signal. In the first case the visual signal was treated symbolically, while in the second it was used to provide subsymbolic information about the corresponding acoustic signal. These two cases are two points on a continuum of speech descriptions. Other

descriptions, such as description of the articulators themselves, could also have been used.

A better understanding of the visual and acoustic sensory systems in humans and other animals will lead to better artificial sensors and their effective integration. Acoustic speech recognition systems, by using models of the human cochlea as a preprocessor, are already benefitting from what is known about the human auditory system. Synthetic cochleas that can process massive amounts of sensory data in real time already have been fabricated in analog VLSI (Mead, 1989). The output of these chips is a highly distilled, parallel and distributed representation of the acoustic signal. Our results are an encouraging first step toward solving the problem of fusing multiple sources of distributed sensory data. Massively-parallel network models could provide the means by which distributed representation can be integrated in real-time for producing rapid recognition and decisive actions for automated systems.

REFERENCES

Bernstein, L.E., Eberhardt, S.P. 1986. Johns Hopkins Lipreading Corpus I-II Johns Hopkins University Baltimore, MD.

Elman, J.L. 1988, Finding structure in time. CRL Technical Report 8801, Center for Research in Language. University of California, San Diego, CA.

Jordan, M.I. 1988. Supervised learning and systems with excess degrees of freedom. COINS Technical Report 88-27, Computer and Information Science, University of Massachusetts at Amherst.

Mead, C. 1989. **Analog VLSI and Neural Systems**, Addison-Wesley: New York, NY

Petajan, E.D. 1987, An improved Automatic Lipreading System To Enhance Speech Recognition. AT&T Bell Laboratories Technical Report No. 11251-871012-111TM. Murray Hill, NJ.

Rumelhart, D.E., McClelland, J.L., & the PDP Research Group, 1986, **Parallel Distributed Processing: Explorations in the Microstructures of Cognition**. The MIT Press: Cambridge, MA 1986

Summerfield, Q. 1987. Some preliminaries to a comprehensive account of audio-visual speech perception in Hearing by Eye: The Psychology of Lip-Reading Eds. B. Dodd and R. Campbell Lawrence Erlbaum Assoc. Publishers: Hillsdale, NJ.

FIGURE CAPTIONS

Figure 1. System used to combine visual and acoustic speech information. A simple vowel recognizer was constructed to receive speech signals from the two modalities. Independent estimates of the vocal tract transfer function were produced and then combined with a weighted average before being passed to the recognizer. A neural network was trained to perform the mapping of the image into the estimated envelope of the acoustic spectra. Noise was introduced into the acoustic speech signal and the improvement due to the visual information was assessed.

Figure 2. Intelligibility of noise-degraded speech as a function of speech-to-noise ratio in dB. The lower curve shows the performance of the recognizer under varying signal-to-noise conditions using only the acoustic channel. The intermediate dashed curve shows the performance when the two independent estimates are equally weighted. The top curve shows the improved performance by using a weighting function based on the signal-to-noise. When the visual signal is used alone, the percent correct is 55% across all S/N levels.

Figure 3. Spectrograms created from the actual acoustic spectra are compared to visually-estimated spectra for the sentence: "We will weigh you". Individual spectral estimates were converted to a grey scale and then aligned by frequency as a function of time. Actual acoustic data from the test set are shown on the left and estimates produced by the feedback neural network model are shown on the right.

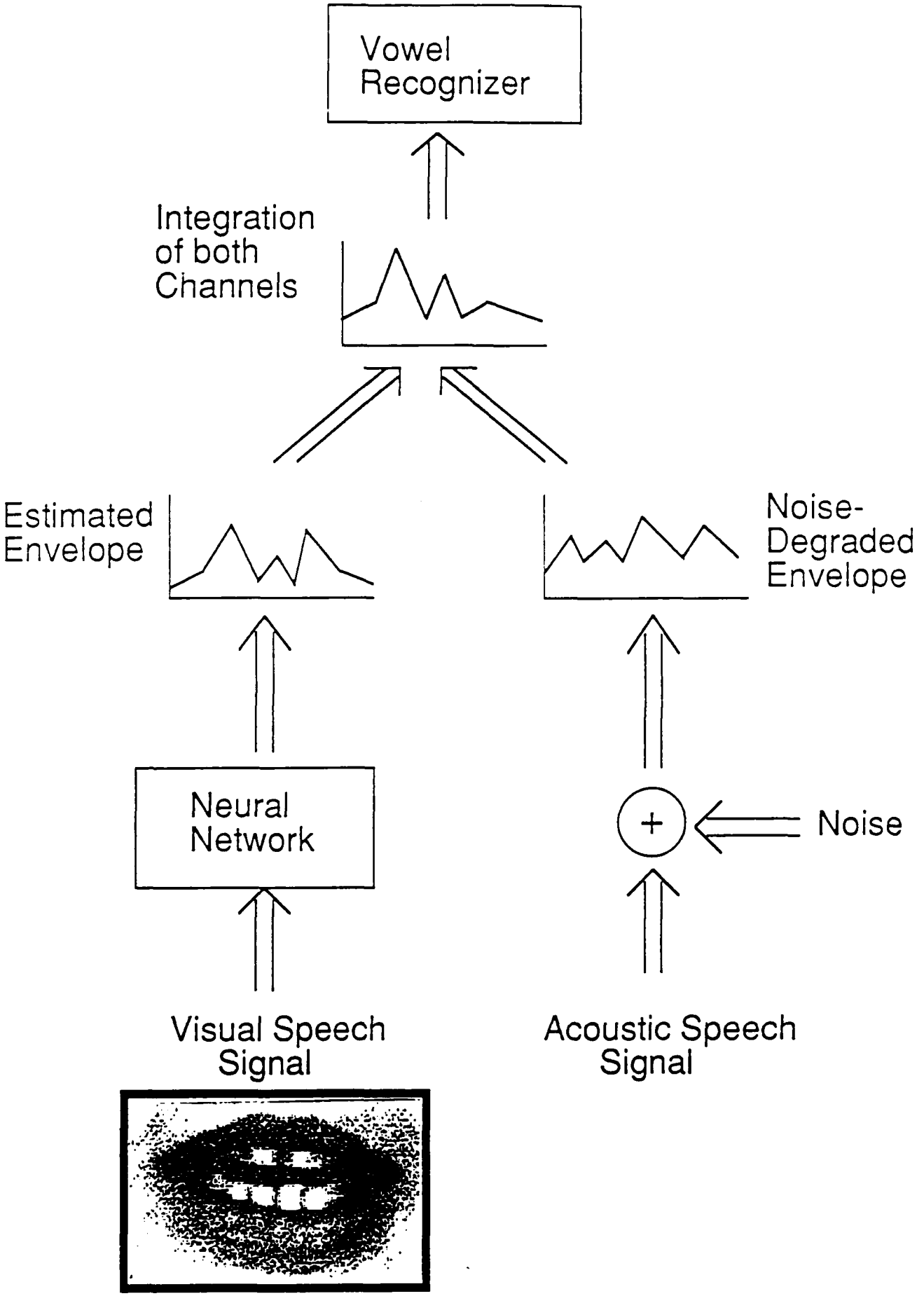
Vowel Recognizer
Integration of both Channels
Estimated Envelope
Noise-Degraded Envelope
Neural Network
+
Noise
Visual Speech Signal
Acoustic Speech Signal

Figure 1

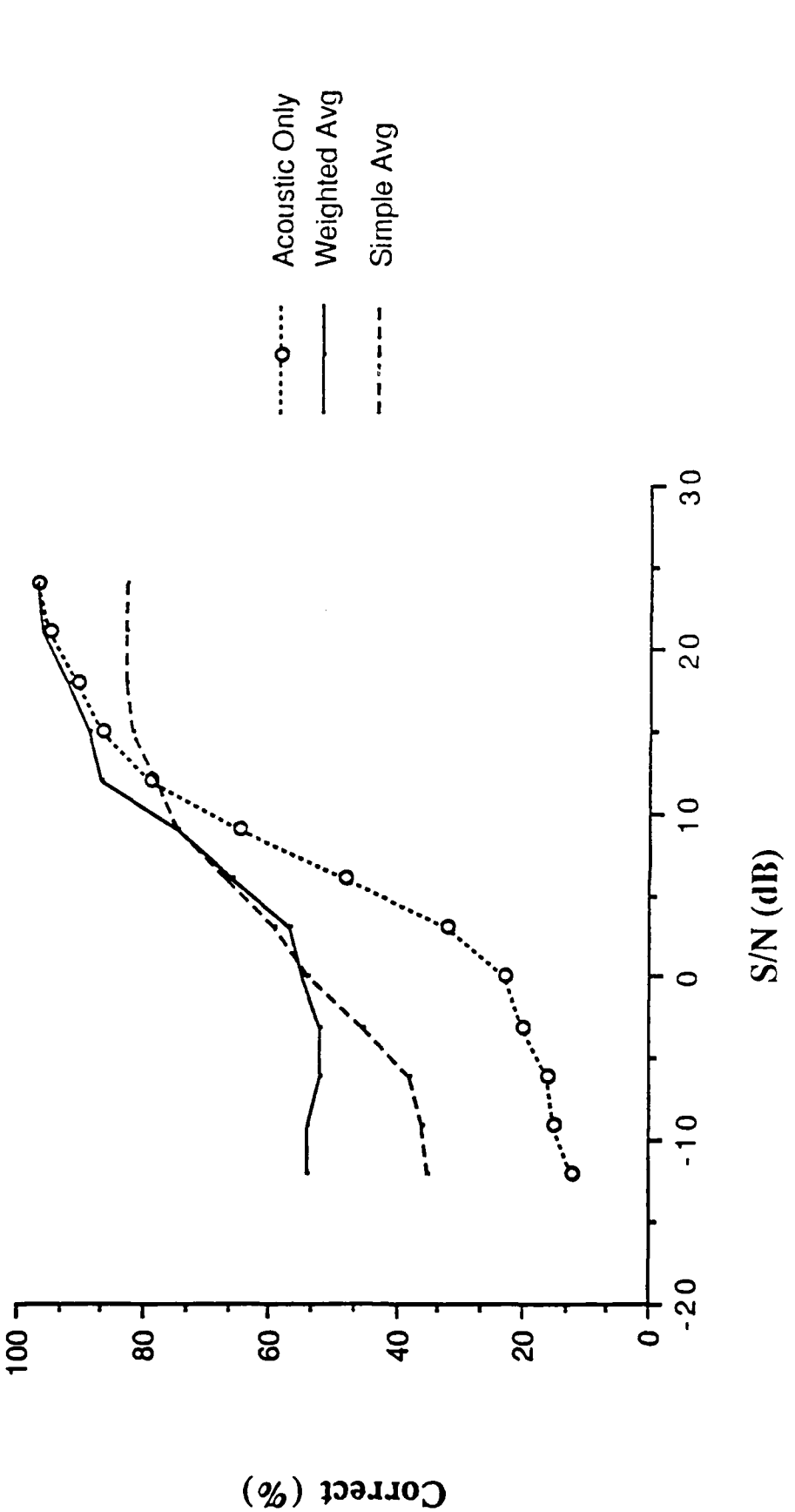
Acoustic Only
Weighted Avg
Simple Avg
Correct (%)
100
80
60
40
20
0
-20
-10
0
10
20
30
S/N (dB)

Figure 2

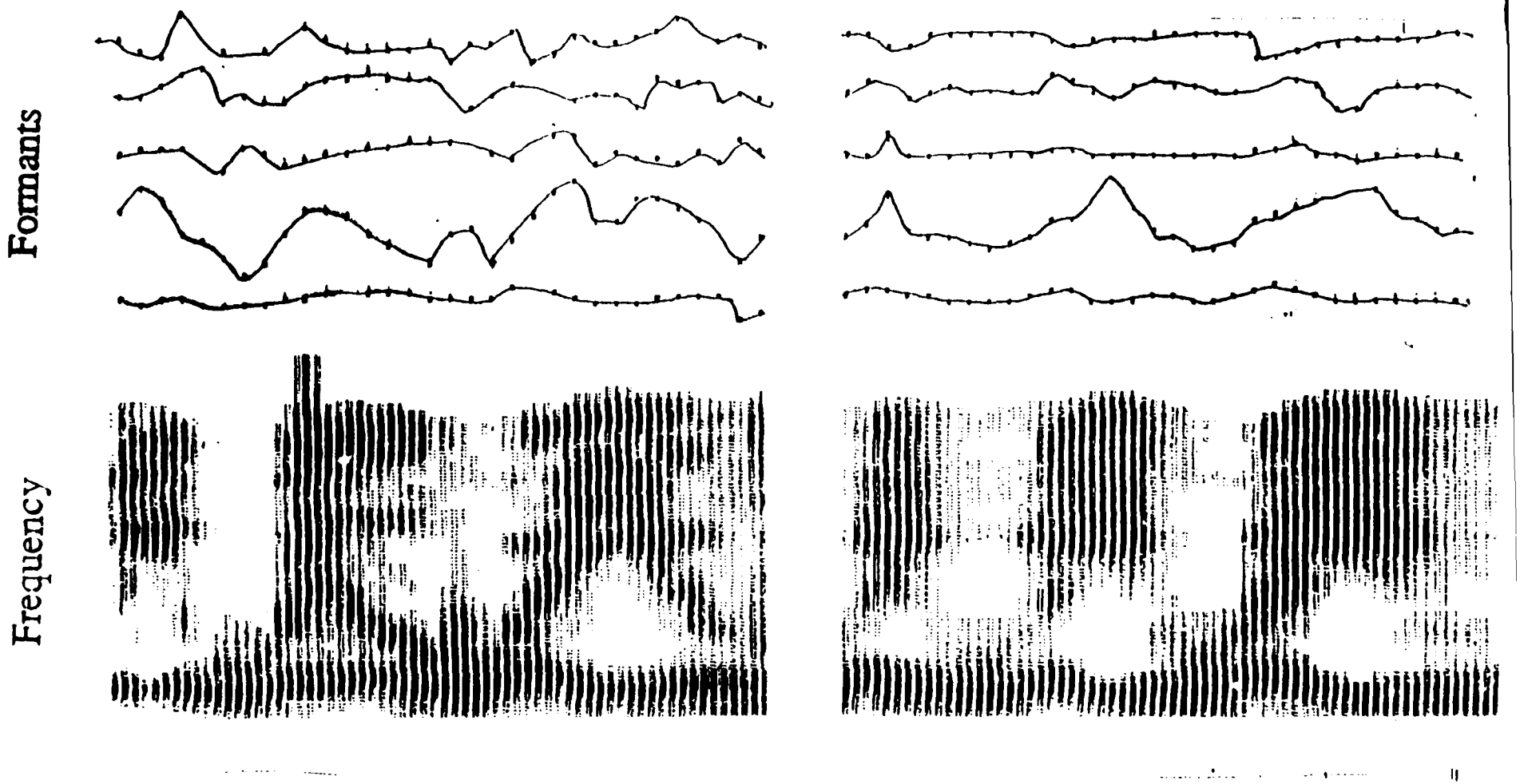
Formants
Frequency
WE WILL WEIGH YOU
WE WILL WEIGH YOU
Acoustic spectra
Estimated spectra

Figure 3